CIAT
International Center for Tropical Agriculture
Since 1967 Science to cultivate change

CGIAR

RESEARCH PROGRAM ON
Climate Change,
Agriculture and
Food Security

CCAFS

Data Quality Assurance

March 2018

Stats4SD

Introduction

Quality control of data is an integral part of all research and takes place at various stages, during data collection, data entry or digitisation, and data checking. It is vital to develop suitable procedures before data collection starts.

Data quality checking is the process of reviewing the data to discover inconsistencies and other anomalies and performing data cleaning activities to improve data quality. Quality control at different stages is part of setting up systems for Data Quality Assurance (DQA).

A dataset is almost never 100% clean, but the aim is to produce datasets that are as error-free as possible. If you have obvious mistakes in your data, you may find you have difficulties justifying your results and conclusions. On the other hand, if you can produce proof that you have systematically checked your data and have eliminated as many errors as possible, then your results will have greater credibility.

In this guide we aim to provide you with hints and suggestions on dealing with anomalies in your data. The exact steps you take will of course vary according to the nature of your data, but the general principles apply to all datasets.

When to carry out Data Checks

You should be aware that errors in data can occur at any stage;

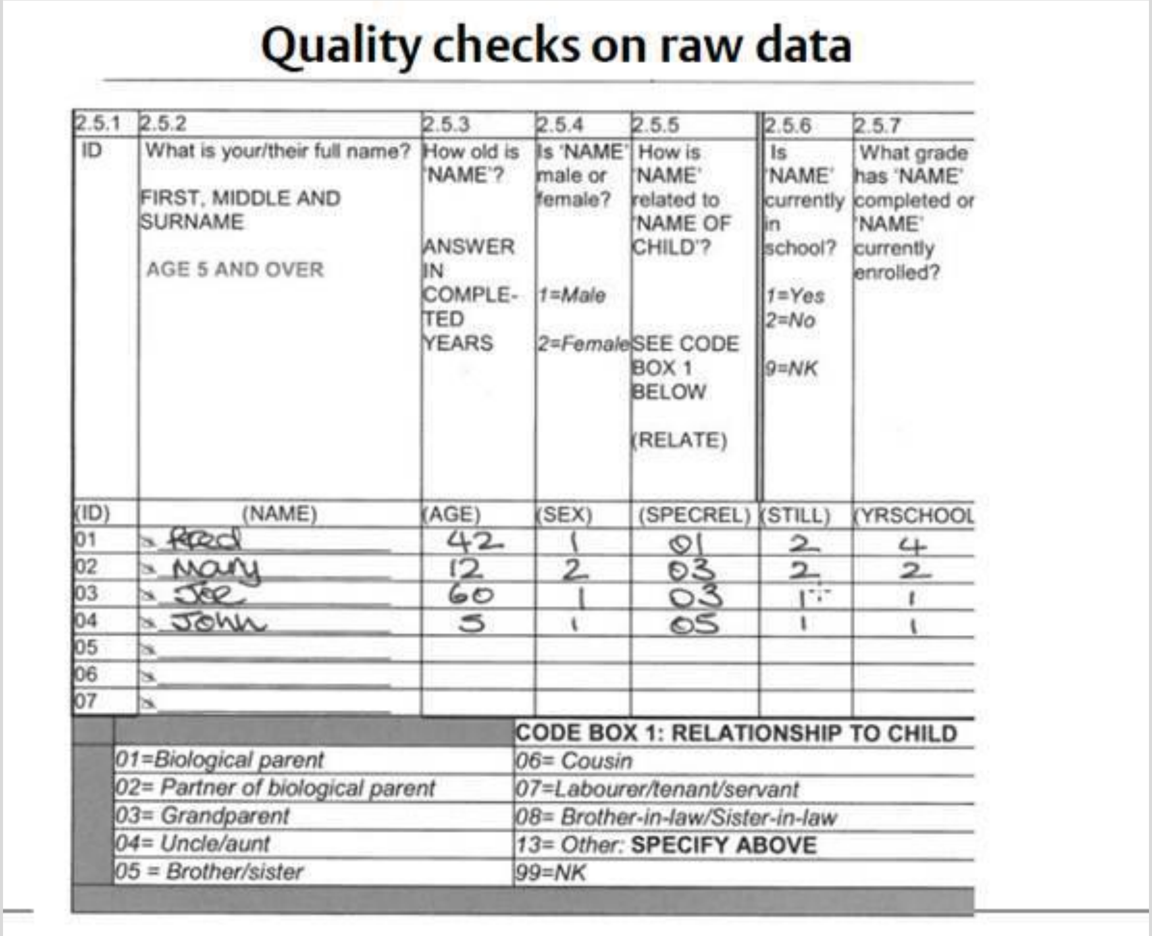
- During the design of your data collection instruments and procedures – strictly speaking there is no actual data at this stage, but you should be aware of the data you wish to collect and have in place mechanisms for minimising errors later on;
- During data collection – e.g. the measuring instrument was not properly calibrated, or it was read incorrectly, or the value was recorded wrongly;
- During data entry – data entry errors are common, for example it is very easy to hit the wrong key or put the decimal point in the wrong place;
- During data manipulation & analysis – e.g. values may be truncated when transferring between different software packages; you may make a mistake in calculation; etc.

Therefore, it is essential that you remain vigilant and check for errors at each stage of the process. Ideally data should be entered and checked for errors as soon as possible after data collection. If this is done in a timely fashion it is often possible to return to the field to check on anomalous values. If data entry is not being done in the field, you should at least do a manual check of your completed data collection instrument – e.g. your experimental collection sheets or survey questionnaires – to make sure there are no obvious errors.

As an example, the image in Figure 1 was taken from a completed questionnaire. The survey selects a child from each household then records information about all household members including their relationship to the selected child. If you look closely at Figure 1 you should notice that the second person in the list (Mary), is recorded as being the grandparent of the selected child (relationship code 03) – however, this is impossible as Mary is only 12yrs old. There is clearly a mistake either in

Mary's age or in the relationship code used. If this error was discovered before leaving the field, it would be relatively easy to revisit the household and find the correct data.

Figure 1 - Household Roster



Quality checks on raw data

2.5.1	2.5.2	2.5.3	2.5.4	2.5.5	2.5.6	2.5.7
ID	What is your/their full name? FIRST, MIDDLE AND SURNAME AGE 5 AND OVER	How old is 'NAME'? ANSWER IN COMPLETED YEARS	Is 'NAME' male or female? 1=Male 2=Female	How is 'NAME' related to 'NAME OF CHILD'? SEE CODE BOX 1 BELOW (RELATE)	Is 'NAME' currently in school? 1=Yes 2=No 9=NK	What grade has 'NAME' completed or 'NAME' currently enrolled?
(ID)	(NAME)	(AGE)	(SEX)	(SPECREL)	(STILL)	(YRSCHOOL
01	Fred	42	1	01	2	4
02	Mary	12	2	03	2	2
03	Joe	60	1	03	1	1
04	John	5	1	05	1	1
05						
06						
07						

	CODE BOX 1: RELATIONSHIP TO CHILD
01=Biological parent	06= Cousin
02= Partner of biological parent	07=Labourer/tenant/servant
03= Grandparent	08= Brother-in-law/Sister-in-law
04= Uncle/aunt	13= Other: SPECIFY ABOVE
05 = Brother/sister	99=NK

Such checks and household revisits should be done whether digital or paper collection is used.

Before Data Collection – Checklists

It is always a good idea to produce lists of things you can check for in your data, and these checklists can be prepared prior to data collection. It's worth making these checklists as comprehensive as possible. For example, for the section of data we can see in Figure 1, we might have the following checks:

- □ Is the age consistent with the relationship? E.g. parents should be at least 12yrs of age, grandparents at least 25yrs, etc.
- □ Is the age consistent with whether or not the person is currently in school? For example, in Figure 1, Joe is 60yrs old but is currently in school – is this feasible – does the question include adult education classes?
- □ Is the age consistent with the education level (completed grade) – e.g. you would not expect a 5yr old to have completed grade 10?
- □ Are the written values all clear and easy to read?

This sort of checklist can be used in the field and is also very useful for adding checks into the data entry system or, in the case of mobile data collection, in the mobile application.

During Data Collection

During data collection there may be a team of enumerators with one or two supervisors working in the field. At the end of each day/session of data collection, the enumerators and supervisors should go through the checklist for each questionnaire or data collection form and highlight any problems with the data. Again, this should be done whether data are being collected digitally or on paper. Although digital data collection systems can be programmed to trap many potential errors, it is not always possible to include checks for every scenario.

During Data Entry

For mobile data collection the data entry phase is combined with the data collection and as already mentioned many checks can be incorporated into the system itself. This is also true when entering data from paper questionnaires. When the data entry system is produced, an accompanying document should be produced detailing the checks and skips that have been incorporated. For example, it should be possible to program the system so that 12yr old grandparents are not accepted. However, there is a limit to what can be included in the system without it becoming too unwieldy. It should be remembered that the data entry system is often produced by a consultant with expert knowledge of producing such systems but without expert knowledge of the type of data being collected. For example, let's assume you were collecting birth weights and wanted the data entry system to prevent non-feasible entries in this field. You would need to clearly define the accepted range of values for this field. Alternatively, you may prefer to check these values after data entry.

For non-survey data, checks for completeness should be done. Check that the number of data rows is as expected; are there any gaps in the data that need to be addressed? If there are missing data, then add comments giving explanations as to why these data are missing.

Double Data Entry

Double data entry is obviously not relevant for mobile data collection. If using paper questionnaires or data collection sheets though we would recommend double-data entry (DDE). This is where the data are entered twice, in separate files or separate copies of the data entry system, and by different individuals. The two data files generated from this process are then compared and any discrepancies are checked against the original data collection sheets or questionnaires. The basic idea behind DDE is that two individuals are unlikely to make the same mistakes on data entry.

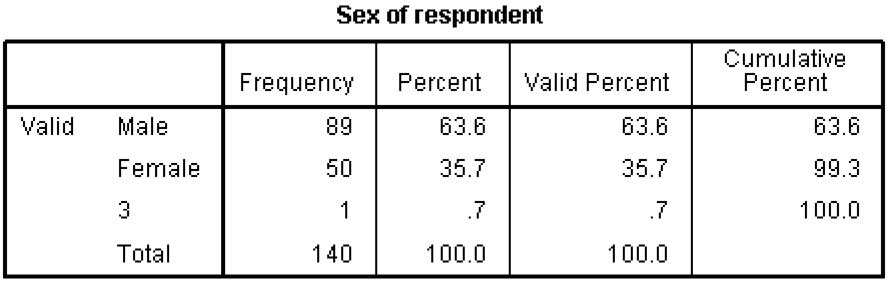
DDE clearly adds time and effort to the data entry process and there is some debate over its value. However, we have yet to be convinced that there is a better or simpler method of data checking available and feel strongly that we should use all the tools at our disposal to help ensure the quality of our data.

Many software packages incorporate tools for DDE. CSPro for example, includes a Data Compare tool which does the comparison of the two data files. The results of the data comparison can then be included in your quality assurance documentation.

After Data Entry

Once the data have been entered you can compare across records looking for extreme cases, and, in the case of categorical data, looking for values outside the range of possible categories. We would recommend running simple frequencies tables for all categorical variables. For example, in Figure 2 you can clearly see there is a value outside the accepted range.

Figure 2 - Category value outside range

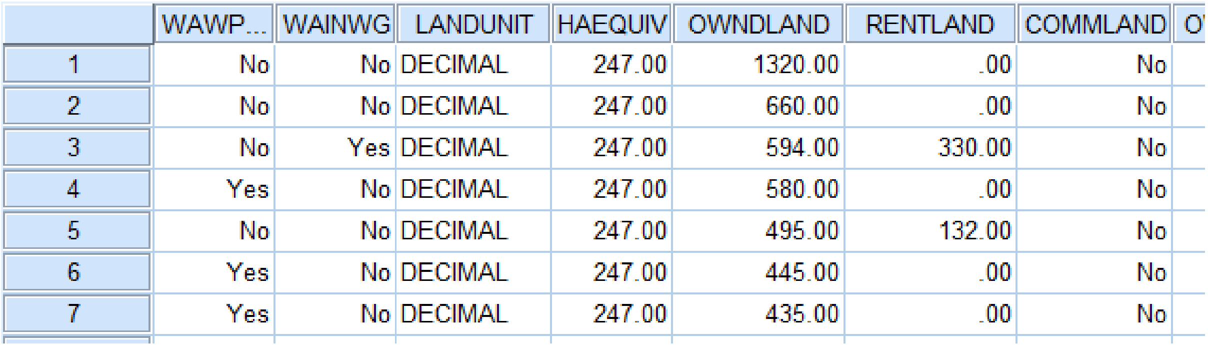


Sex of respondent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	89	63.6	63.6	63.6
	Female	50	35.7	35.7	99.3
	3	1	.7	.7	100.0
	Total	140	100.0	100.0	

For continuous variables, such as area of land, we suggest producing simple descriptive statistics to see the range of values. Sorting the data is also worthwhile. In Figure 3 the data have been sorted by area of land owned and you can clearly see that the first household in the list appears to own twice as much land as any other household.

Figure 3 - Sorted data



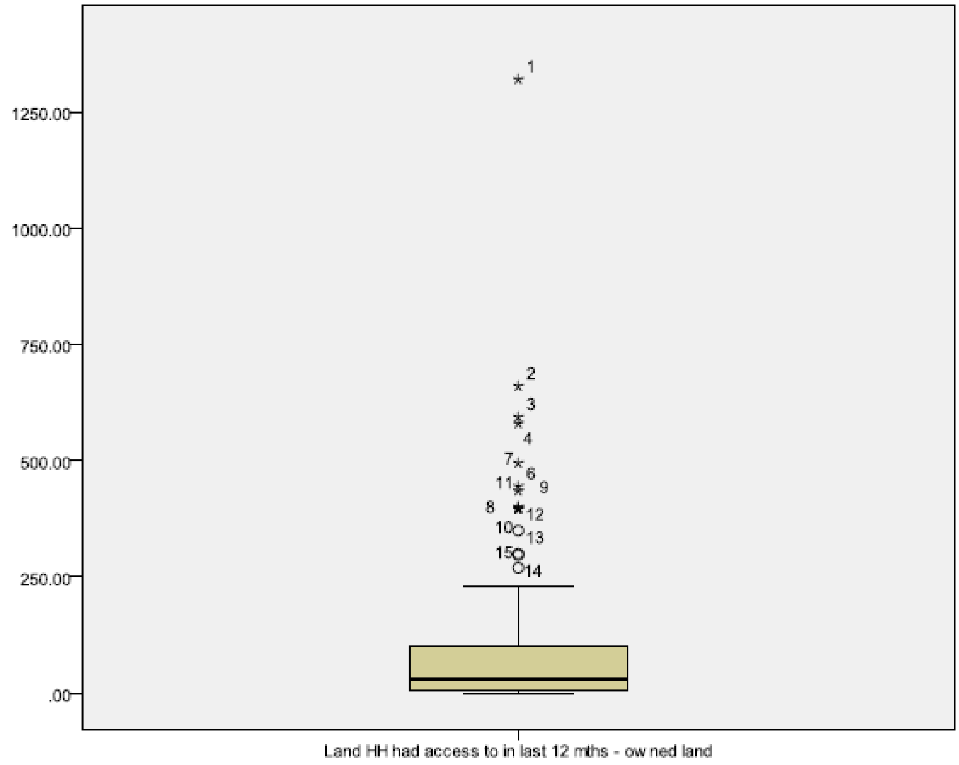
	WAWP...	WAINWG	LANDUNIT	HAEQUIV	OWNDLAND	RENTLAND	COMMLAND	O
1	No	No	DECIMAL	247.00	1320.00	.00	No	
2	No	No	DECIMAL	247.00	660.00	.00	No	
3	No	Yes	DECIMAL	247.00	594.00	330.00	No	
4	Yes	No	DECIMAL	247.00	580.00	.00	No	
5	No	No	DECIMAL	247.00	495.00	132.00	No	
6	Yes	No	DECIMAL	247.00	445.00	.00	No	
7	Yes	No	DECIMAL	247.00	435.00	.00	No	

Dealing with Outliers

The first task when it comes to outliers is to identify them. The data from Figure 3 have been plotted using a boxplot[1] (Figure 4) and this clearly shows the outlier. Boxplots are a useful tool to help you visually assess the spread of the data and whether it clusters in the regions that would be expected. But is the outlier we see in Figure 4 an error in the data or is this just a particularly wealthy farmer with a large land-holding? If it is possible to return to the field and check the data then that would be the best options, but this may not be possible.

[1] In descriptive statistics, a boxplot (also known as a box-and-whisker plot) is a convenient way of graphically depicting groups of numerical data through their five-number summaries: the smallest observation, lower quartile, median, upper quartile and largest observation. A boxplot may also indicate which observations, if any, might be considered outliers.

Figure 4 - Outlier shown in boxplot



Often you need to make a decision based on logical deduction. Is the value within the natural range of the measurement? If so, the decision would verge towards keeping the observation in the dataset. If not, then you may be justified in removing it. Whatever decision you make, it should be documented along with the reasons behind your decision.

Transfer of Data from Field to Base

As part of your data management plan and your data quality control, you should document how you intend transferring your data from the field to the office. This might involve the transport of paper questionnaires or collection sheets, or it might involve the method you intend to use to transfer data from hand-held devices. This includes the frequency of transfer and any systems you have in place for ensuring safe delivery.

Electronic Transfer

When digital data collection is used, you will need a method of transferring and checking the data. As files are transferred you need to ensure they arrive uncorrupted – there should be mechanisms for checking the file(s) on arrival – don't delete the file(s) from the hand-held device until successful transfer has been confirmed.

Do you have a backup system for your hand-held device? How will you deal with updates being sent and how will you deal with duplicates?

Paper Questionnaires

For paper questionnaires we recommend including a “Data Handlers” section on the front sheet with space for the enumerator and supervisor to sign and date once they have carried out visual consistency checks.

Back in the office you should have some sort of logging process. At the simplest level you could just record the number of questionnaires or collection sheets sent out and match this against the number received. Alternatively, you might consider numbering the sheets before going to the field and keeping a log of sheets as they are returned. This way you will have more of an idea of which sheets/questionnaires are missing.

Whether transferring data electronically or on paper, you will need a system of checks and balances to confirm that:

- Everything sent from the field is received in the office;
- Everything received is uncorrupted and is complete;
- The expected number of questionnaires/recording sheets are received.

Backups

Backing up your data is an important aspect of data management and data quality control. When making data corrections for example, it is easy to make mistakes and accidentally make the correction to the wrong record or variable. It is therefore essential that you have backup versions that you can return to if necessary. Some types of cloud storage such as Dropbox, automatically keep previous versions of files that you can return to. However, it is best not to rely on others to do your backups for you.

Transition from Raw to Primary Data

Raw data encompasses everything that was gathered in the data collection process. This might include information about who collected the data, dates of collection and checking, details of the device used for data collection, etc. When collecting data digitally the resulting data files often include notes and/or the results of intermediate calculations. Some of this raw data will need processing further before any analysis can be done. The result of this processing is the primary data.

This data management pack includes a separate document describing the transition from raw to primary data.

Audit Trail

We would strongly recommend keeping an Audit Trail for your data. In other words, a document detailing the checks run on the data and any corrections/changes made. Decisions on how you dealt with any outliers should be included in this same document. If you have used double data entry, then the results from the data comparisons can also be included in the Audit Trail. The Audit Trail can form part of the data management report and/or the data quality report.

Units of Measurement

A common problem when collecting data, particularly through questionnaires, is to do with inconsistencies in the unit of measurement used. Often the unit of measurement is specified on the data collection sheet or in the questionnaire. For example, the data in Figure 5 are taken from a child nutrition study in which children were weighed at the start of the study and again two weeks later. The data collection sheet indicates that weight should be in kilograms. However, there is clearly something odd with child number 002 who appears to have more than doubled his/her weight in the two weeks. When this value was investigated it was found that in this instance the value had been recorded in pounds. When the unit of measurement is specified on the instrument you must ensure you use that unit when collecting and recording your data.

Figure 5 - Weight in kgs

Child ID	Weight before (Kg)	Weight after (Kg)
001	16.3	17.1
002	17.2	39
003	17.4	17.5
004	16.5	17.2

On the other hand, you may find the situation where the unit of measurement is not fixed. In this case you should record the unit used in a separate variable along with the conversion factor to convert the data to a standard unit. For example, in Figure 6 the area of land owned by farmers was recorded, but this was specified in the local unit. The local unit was recorded along with the conversion factor to convert these values to hectares. Thus

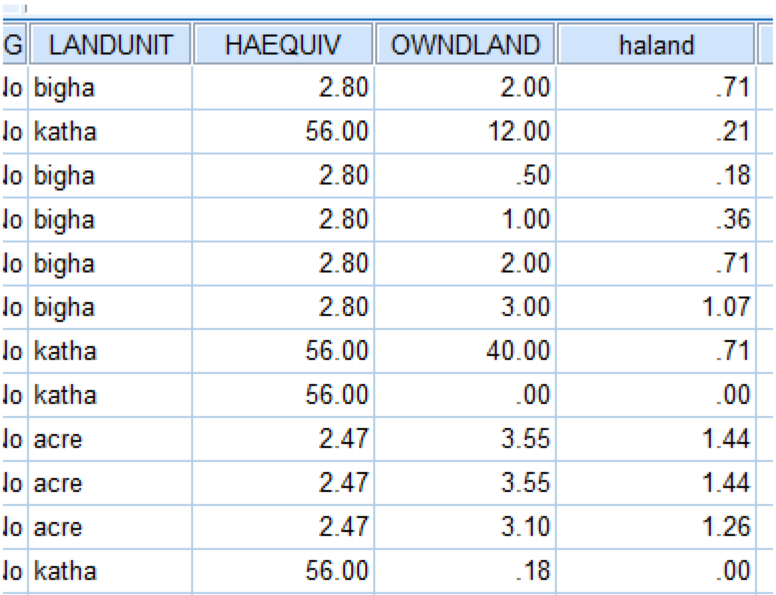
1	hectare	=	2.47	acres
1	hectare	=	2.80	bigha
1	hectare	=	56	katha

The data values are then recorded in the local units – field staff should not do hand calculations but instead you should produce syntax files in your chosen statistics package to do the conversion for you. In this example the column "haland" has been calculated by the statistical software and is the equivalent areas of land in hectares. The formula used was:

haland = OWNDLAND / HAEQUIV

It is important in any analysis for "haland" to be used rather than "OWNDLAND". This conversion is part of the process of transferring from raw to primary data.

Figure 6 - Different units of measurement



G	LANDUNIT	HAEQUIV	OWNDLAND	haland
lo	bigha	2.80	2.00	.71
lo	katha	56.00	12.00	.21
lo	bigha	2.80	.50	.18
lo	bigha	2.80	1.00	.36
lo	bigha	2.80	2.00	.71
lo	bigha	2.80	3.00	1.07
lo	katha	56.00	40.00	.71
lo	katha	56.00	.00	.00
lo	acre	2.47	3.55	1.44
lo	acre	2.47	3.55	1.44
lo	acre	2.47	3.10	1.26
lo	katha	56.00	.18	.00

Local Knowledge

In the above example it is important that hand calculations are not done and that the original measurements are entered along with the correct conversion factor. As far as the conversion factor is concerned this must be in the "right direction" – for example we are recording the number of local units in one hectare as the conversion factor. This must be made clear to the enumerators otherwise some may record the number of hectares in one local unit. This happened in one of the CCAFS sites for the household baseline survey where the conversion factor for acres was recorded as 0.40 rather than 2.47. Consequently, when the new variable was calculated, farmers appeared to have much more land that was actually the case. For example, the farmer in Figure 6 with 3.1 acres of land would have been calculated as having 7.75 hectares (3.1/0.4) of land rather than the correct 1.26 hectares.

This is where local knowledge can be very useful – someone familiar with the region would be able to tell you that farmers do not own that much land. When your data appears to contradict local knowledge, you should check your data.

Versioning

During the lifetime of a research project there are likely to be different versions of the data as errors are identified and corrected. Some researchers like to keep different versions particularly if they have started analysis before all the errors are identified. If you are going to keep previous versions, you should develop a naming convention for your data files together with a log outlining the differences between the versions. Your naming convention might just be to include the version number in the filename, or you might prefer to include the date in the filename. Either way, the method you use should be documented and agreed by parties and everyone involved in the project should be aware of which version is the latest or definitive version and when new versions become available.

Summary

We have seen that errors can creep into your data at any stage of the project lifecycle. Data checking is not just something to be done between data entry and analysis but continues throughout the project. A dataset is rarely 100% error-free. Therefore:

- Be vigilant – if something looks odd in your data or is unexpected, check it;
- Keep an Audit Trail detailing what checks and changes you make;
- Use local knowledge to check what the data appears to be telling you;
- If someone tells you "the data have been cleaned" ask to see their Audit Trail.

Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at https://www.youtube.com/channel/UCs7EU95YMjhvNozJKCD92xQ/playlists. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on Data Quality Checking which is available at the following link:
https://www.youtube.com/watch?v=vbxvtIbqkPA&index=15&list=PLK5PktXR1tmNRaUPsFiYlyhg2lui0xgpj